UPPSALA
UNIVERSITET

Data Lab 2025

QC in PLINK2 | GWAS | Meta-Analysis

PLINK2.0 alpha

- A new version of PLINK: a free, open-source whole genome association analysis tool
  - Association tests (GWAS), genetic data preprocessing, post-GWAS processing
  - For GWAS: supports quantitative (continuous) and dichotomized qualitative (binary) traits
  - Not for genotype calling or result visualization
  - Not taking genetic relatedness into account: need to remove related individuals if any
- Available on UPPMAX HPC
- More reference: https://www.cog-genomics.org/plink/2.0/

UPPMAX HPC

- High-Performance Computing: using supercomputers and computer clusters to solve advanced computational problems.
- UPPMAX: The computer center of Uppsala University, part of NAISS (The National Academic Infrastructure of Supercomputing in Sweden)
- Three clusters:
  - Bianca: sensitive data, e.g., real patient data, not connected to internet
  - **Rackham:** for local projects with PIs at UU, and for courses
  - Snowy: for courses and some local projects
- https://docs.uppmax.uu.se/

Log in to Rackham/Snowy

- Most convenient: From the laptop's terminal with SSH:
  - `ssh <username>@rackham.uppmax.uu.se`
  - Enter your password
  - Use regular Linux commands
  - The lab materials: `/crex/proj/uppmax2024-2-1/DATA_LAB/data`
- Website: ThinLinc (https://rackham-gui.uppmax.uu.se/)
  - Need to set up a two-factor authentication
  - Not necessary
- https://docs.uppmax.uu.se/getting_started/login_rackham/
- https://docs.uppmax.uu.se/cluster_guides/snowy/
- Download results from Rackham:
  - `sftp <username>@rackham.uppmax.uu.se`
  - `get <filename>` (assumed it is at your home directory: /home/<username>)
  - To upload: `put <filename>`
  - To logout: `exit`

UPPSALA UNIVERSITET

Handling big datafiles and jobs

- When you log in, you are at the “login node”:
  - It is shared
  - Only do short and light things: editing, deleting, moving files
    - Start interactive session
    - Schedule jobs (sbatch)
- Jobs
  - Performing computational jobs with large datafiles and abundant codes
  - Developing computational scripts with many line, for large datafiles
- Interactive session (write command lines one-by-one)
  - `interactive -A uppmax2024-2-1 -M snowy -t 2:00:00 -n 2`
    - Using 2 cores (no more than 16) for 2 hours
  - When you finish, type `exit` to quit the interactive nodes
- Job scheduler (when you have prepared a shell script, and when your job will take hours)
  - `sbatch -M snowy -A uppmax2024-2-1 your_analysis.sh`

UPPSALA UNIVERSITET

Job scheduler

- In your .sh script:

```
#! /bin/bash -l
#SBATCH -A <project-number>
#SBATCH -p core
#SBATCH -n 2
#SBATCH -t 3:00:00
#SBATCH -J <your-job-name>
#SBATCH --mail-type=ALL
#SBATCH --mail-user=<your-uppmax-account-email>

module load <packages-you-need>

some_path_variable="/crex/proj/uppmax2024-2-1/DATA_LAB/data"

plink2 <analysis-you-will-be-running>
Rscript <an-R-script.R>
```

You can then access by ${some_path_variable} in following lines

Example plink commands in later slides

- You can write this offline and paste to your editor on Rackham:
  - `nano your-analysis.sh`
  - When you finish, press keys ‘ctrl’ and ‘o’, then ‘Enter’; and ‘ctrl’ and ‘x’

UPPSALA UNIVERSITET

Software for Lab on Rackham

- PLINK2:
  - `module load bioinfo-tools`
  - `module load plink2`
  - Note: If you are running against a binary phenotype, don't use this latest version on UPPMAX: plink2/2.00-alpha-5-20230923
  - Instead: `module load plink2/2.00-alpha-3.7-20221024`
  - There were fatal bugs for binary phenotype in that version
  - Check version: `plink2 --version`
- R:
  - `module load R_packages/4.2.1`
  - Type `R` to enter the coding interface

Data

- Location: `data_loc="/crex/proj/uppmax2024-2-1/DATA_LAB/data"`
  - You will be able to use a shortcut ${data_loc} after typing the line above
  - What's inside the directory: `ls ${data_loc}` Note: if your filename contains space, must add quotes "${data_loc}"
  - PLINK1 files for a quick view of how genotype files look like: .ped and .map files in tab- or space-delimited formats, could be visualized by text editors, or simply:
    - `head ${data_loc}/test.ped | column -t` To make it visually pleasing by aligning the columns in view
    - You can see the number of lines or rows by:
      - `awk 'END {print NR}' ${data_loc}/test.ped` `awk` is a very efficient tool for checking text-files
  - .ped: individual genotypes; .map: description of SNPs.
    - They are headerless (no column names).
- PLINK2 genotypes:
  - .psam: sample information (Family ID; Individual ID; Sex); equivalent of .fam in PLINK1.9
  - .pvar: variant information (SNP positions, alleles); equivalent of .bim in PLINK1.9
  - .pgen: genotype data (compressed, efficient); equivalent of binary format .bed in PLINK1.9
  - epihealth.psam, epihealth.pvar, epihealth.pgen

UPPSALA UNIVERSITET

Quality control (QC)

- Prior to an association analysis, the genetic data has to be checked and filtered to avoid errors or poor quality
  - "Data quality control in genetic case-control association studies." (Anderson et al. 2010, *Nature Protocols* 5: 1564-73)
- QC criteria
  - Missing genotype rates
    - --geno [max per-variant]: filter out variants with missing call rates exceeding the threshold
    - --mind [max per-sample]: does the same for samples
    - Individual call rate: --mind 0.05 removes individuals with >5% missing genotypes
    - `plink2 --pfile ${data_loc}/test **--mind 0.05** --make-pgen --out filtered_file`

      Note: if all input files have the same suffix, i.e., "test", you can use the flag --pfile, otherwise you need to specify each by --pgen, --psam, --pvar; --make-pgen means to generate the filtered files; --out you can specify the output suffix
    - SNP call rate "--geno": sometimes the genotyping of a single SNP is not working properly, such as primer problems; filter SNPs with low genotyping rate from the dataset: --geno 0.05 removes SNPs with >5% missing genotypes.
    - Just add flags to the PLINK2 command line

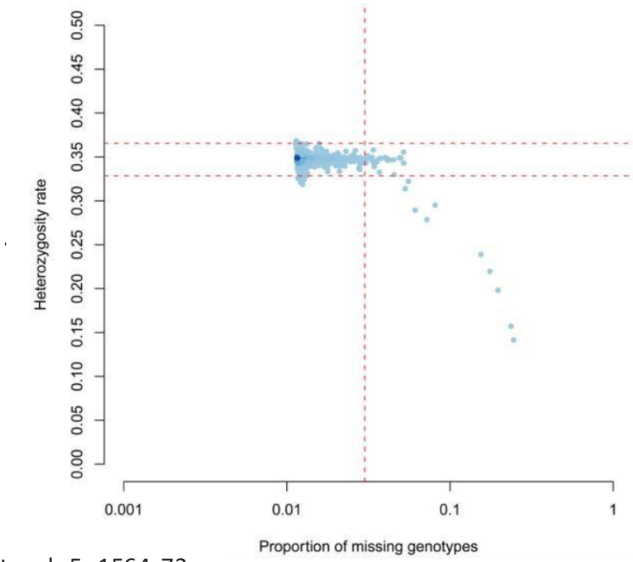
UPPSALA UNIVERSITET

Quality control (QC) cont.

- QC criteria (cont.)
  - Minor allele frequency (MAF)
    - --maf [minimum freq]: filters out variants with allele frequency below the threshold
    - MAF≥0.05 Common variants, for large GWAS, well-powered studies, sample size ~5000+
    - MAF≥0.01 Low frequency variants, for mixed GWAS, disease risk analyses, sample size ~10,000+
    - MAF≥0.001 Rare variants, for sequencing studies, burden tests, sample size ~100,000+
    - **`--maf 0.05`**
  - Hardy-Weinberg equilibrium (HWE)
    - --hwe [p]: filters out variants which have HWE exact test p-value below the threshold
    - `--hwe 1e-4`; `--hwe 0.0001`

Additional steps for raw datasets

- "Inbreeding"
  - --het
  - Computes observed and expected homozygous/heterozygous genotype counts for each sample, reports method-of-moments F coefficient estimates
    - 1-(<observed het. count> / <expected het. count>)
    - Low degree of heterozygosity is a sign of bad genotyping
- Sex imputation
  - --check-sex
  - Checks if sex assignments corresponds to Chr X genotypes
  - Genetic males cannot be heterozygous
- Duplicate variants
  - --rm-dup
  - removes all but one instance of each duplicated-ID variant
- Duplicated or related individuals
  - --make-king: KING-robust kinship estimator
  - Duplicated samples: kinship 0.5
  - 1st-degree: ~0.25; 2nd-degree: ~0.125



Anderson et al. 2010, Nature Protocols 5: 1564-73

UPPSALA UNIVERSITET

Additional steps for raw datasets

- Divergent ancestry
  - Check principal components (PCs); --pca
  - E.g. in UK Biobank

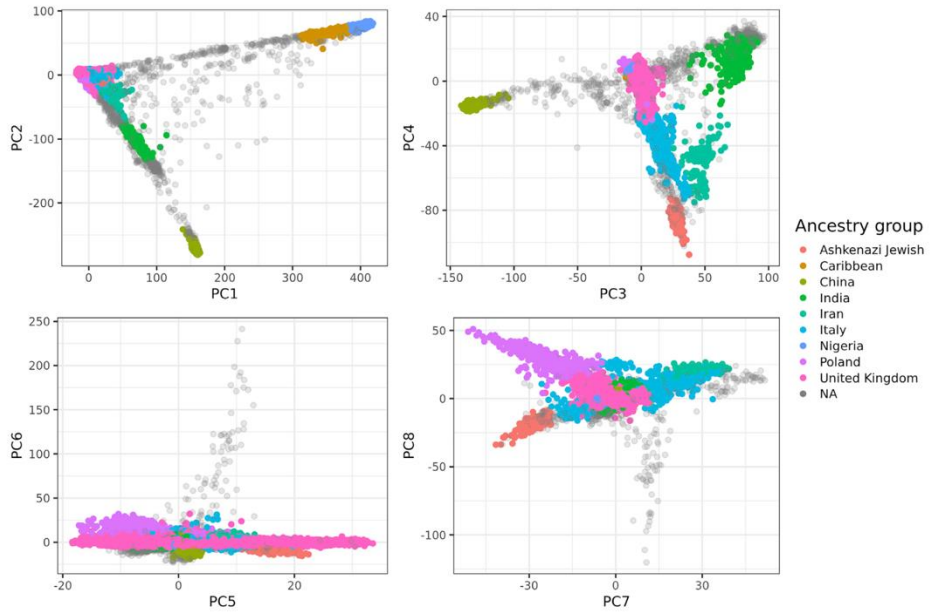


Figure 1. The first eight PC scores of the UK Biobank (field 22009) colored by the homogeneous ancestry group we infer for these individuals
Only 50,000 individuals are represented at random. "NA" means that the corresponding individual is not categorized in any of the nine ancestry groups.

Privé et al. 2022, The American Journal of Human Genetics 109: 12-23

UPPSALA UNIVERSITET

Other useful commands

- --extract <one-column text-file containing the SNP IDs with specific interest>
- --keep <text-file containing individual IDs to be tested>
  - First line: #FID #IID; tab-delimited format
- --export: creates new fileset, after sample/variant filters have been applied, in the formats such as:
  - A: [--export A] Sample-major additive (0/1/2) coding, suitable for loading from R. By default, REF alleles are now counted.
    - This can be adjusted with --export-allele <one-column file with A1 specified>
  - AD: sample-major additive (0/1/2) + dominant (het=1/hom=0)coding
- --make-bed: gives binary file of PLINK1 format.
- In the shell script, you can break lines instead of having everything in one long line as

```
plink2 \
  --pfile … \
  --maf … \
  --out … \
```

UPPSALA UNIVERSITET

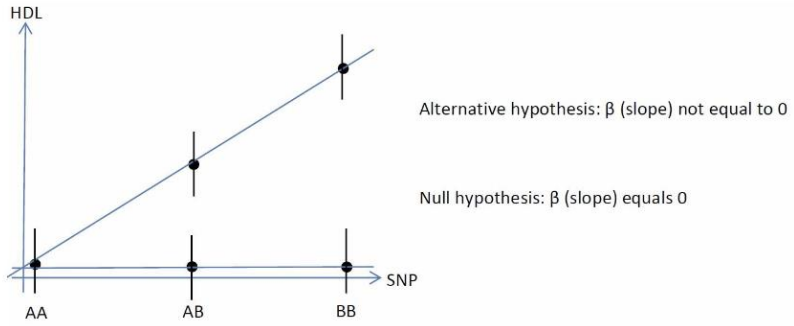
Instructions for the lab:

https://github.com/YADengUU/Data_Lab_2025

UPPSALA
UNIVERSITET

Association test

- Identify the SNPs associated with HDL levels
- Input data:
  - QC-ed Genotype data created in Lab 1
    - Flags to read files: --pfile (PLINK2); --bfile (PLINK1 binary)
    - If the 3 files don't share the same name, specify by pgen/psam/pvar; bed/bim/fam
  - Phenotype file: epihealth_hdl.pheno
  - Covariate file: epihealth_pcs.covar



Preview the phenotype and covariates to see the names of phenotype and covariates:

```
head ${data_loc}/epihealth_hdl.pheno |
column -t
```

Is it a binary or continuous phenotype?

```
head ${data_loc}/epihealth_pcs.covar |
column -t
```

UPPSALA UNIVERSITET

Association test

- --glm hide-covar cols=+a1freq,+ax,+beta
  - --glm is the flag to perform association analysis
  - In the options behind you choose what to show in your results
- --pheno <phenotype-path> Path = directory/filename
- --pheno-name <name-of-phenotype>
- --covar <covariate-path>
- --covar-name <list the covariates, separated by a space in between>
- --out <your customized output name>
- Altogether
  - plink2 --glm hide-covar … --pfile … --pheno … --pheno-name … --covar … --covar-name … --out …

Trick: you don't need the separately created genotype data, the QC flags can be included together in the PLINK2 command line

UPPSALA UNIVERSITET

Read the results

- As mentioned, awk is a highly efficient tool to check data saved as space/tab-delimited form
  - `awk '$N<5e-8' gwas-result-filename`
  - Replace N with the actual column number containing the p-values
  - If you want to save the quick result: add " `> quick-result-name`" at the end
  - To sort in ascending order, such as for the 3rd-column:
    - `awk '$N<5e-8' gwas-result-filename | sort -gkN > quick-result-name`
    - -gkN: numerically (-g) by the N-th column (-kN)
- Read and visualize data In R:
  - `gwas_result <- read.table("gwas-result", header=TRUE)`
  - `dim(gwas_result)` # check how many rows and columns, if not correct, need to modify sep="\t" or sep=" " in read.table().
  - `library(qqman)` Package "qqman" for visualizing GWAS results

```
png(file="Manhattan-plot.png", width=80, height=50, units="mm", res=300)
manhattan(gwas_result, chr="column-of-chrom-num",bp="column-of-SNP-position", p="column-of-p-val", snp="column-of-SNP-ID")
dev.off()
```

  - Q-Q plot: `qq(gwas_result[, "column-of-p-val"])`

Adjust the figure size in png() before plotting; alternatively, if you have saved the GWAS results in your local computer, you can directly plot without png() or dev.off() and see your figures in RStudio

UPPSALA UNIVERSITET

Meta-analysis of GWAS

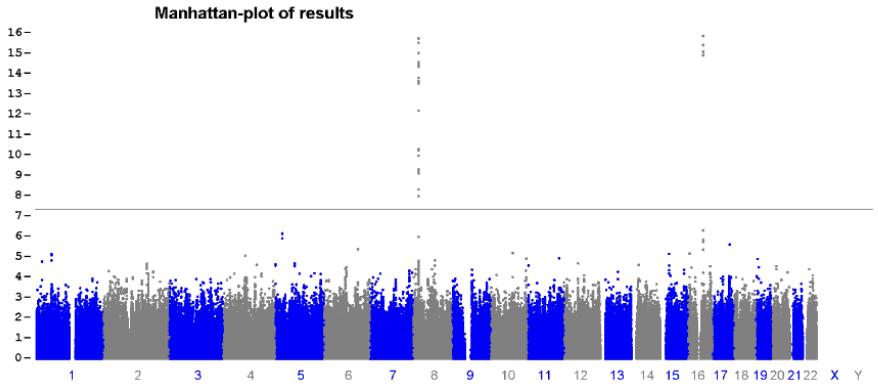
- Most GWAS identify SNPs with small effect sizes and a larger number of statistical tests are performed
- The statistical power to detect a true association between outcome and SNP is low if the sample size is small
- To increase the statistical power, GWAS results from multiple studies are typically combined using meta-analysis

Meta-analysis in METAL

- In this data lab, the regression results from two studies will be combined using a fixed-effects, inverse-variance weighted meta-analysis
- See instructions for the lab (https://github.com/YADengUU/Data_Lab_2025) for:
  - Create a METAL script file, which is a sample.txt file, using text-editor, not Word.
  - Run the METAL program

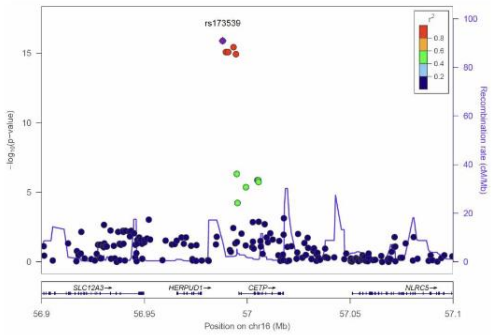
Plot meta-analysis results genome-wide

- R package qqman
- Input: meta-analysis results file with a chr and position column added



Or for a selected region

- Use web-based tool Locuszoom
- Meta-analysis results for a selected region around the CETP gene will be used



UPPSALA UNIVERSITET